Data Mining Capstone
Task 2

I used the 30 most reviewed cuisines for all subtasks to facilitate ease of comprehension and comparison. A larger number of reviews per cuisine would also help in learning better models. I used python scripts (modified versions of the scripts provided, especially adding idempotency to them) to generate the similarity matrix and corrplot from R to plot the visulizations. Each of the following pages has a plot along with the associated analysis of the plot.

The best results were obtained by using the TfIdfVectorizer from scikit-learn with idf weighting and training the matrix obtained on an Lda topic model from the gensim package for 2000 iterations and 5 passes of the dataset. The number of topics chosen for the Lda model was 20. I didn't try to fine tune the Lda parameters any further as these seemed to be producing sensible results.

Task 2.1 Visualization of the Cusine Map

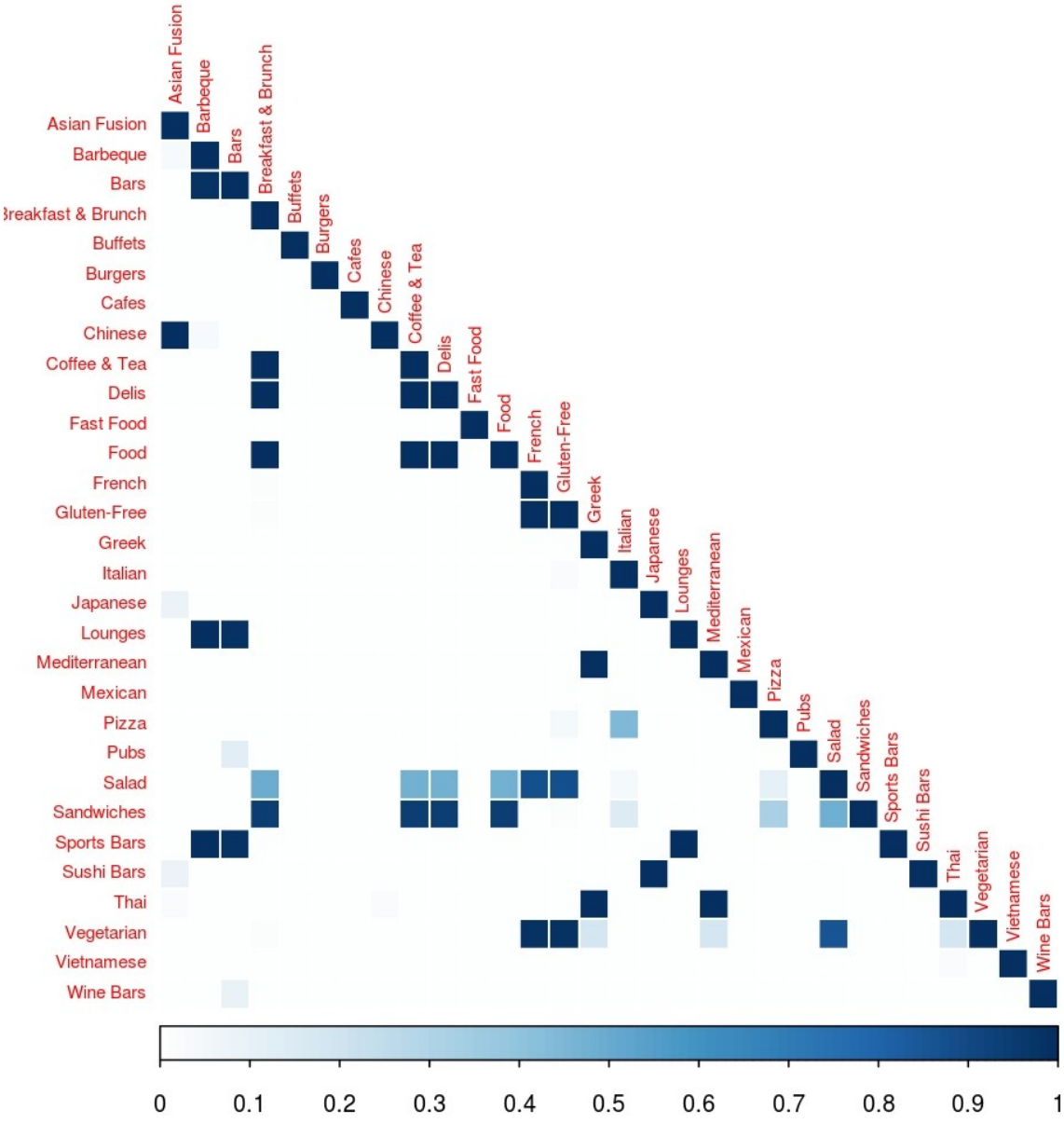

Fig. Cuisine Map obtained with CountVectorizer from scikit-learn.

Analysis: Since CountVectorizer just converts text documents into a collection of token counts, we observe only strong correlations between cuisines picked (like between Salad and Vegetarian, Delis and Breakfast, Chinese and Asian Fusion). The strong correlations must have been because of the same words occuring in the reviews for these cuisines. I have added clustering to the same cuisine map to better visualize which cusines appear similar.

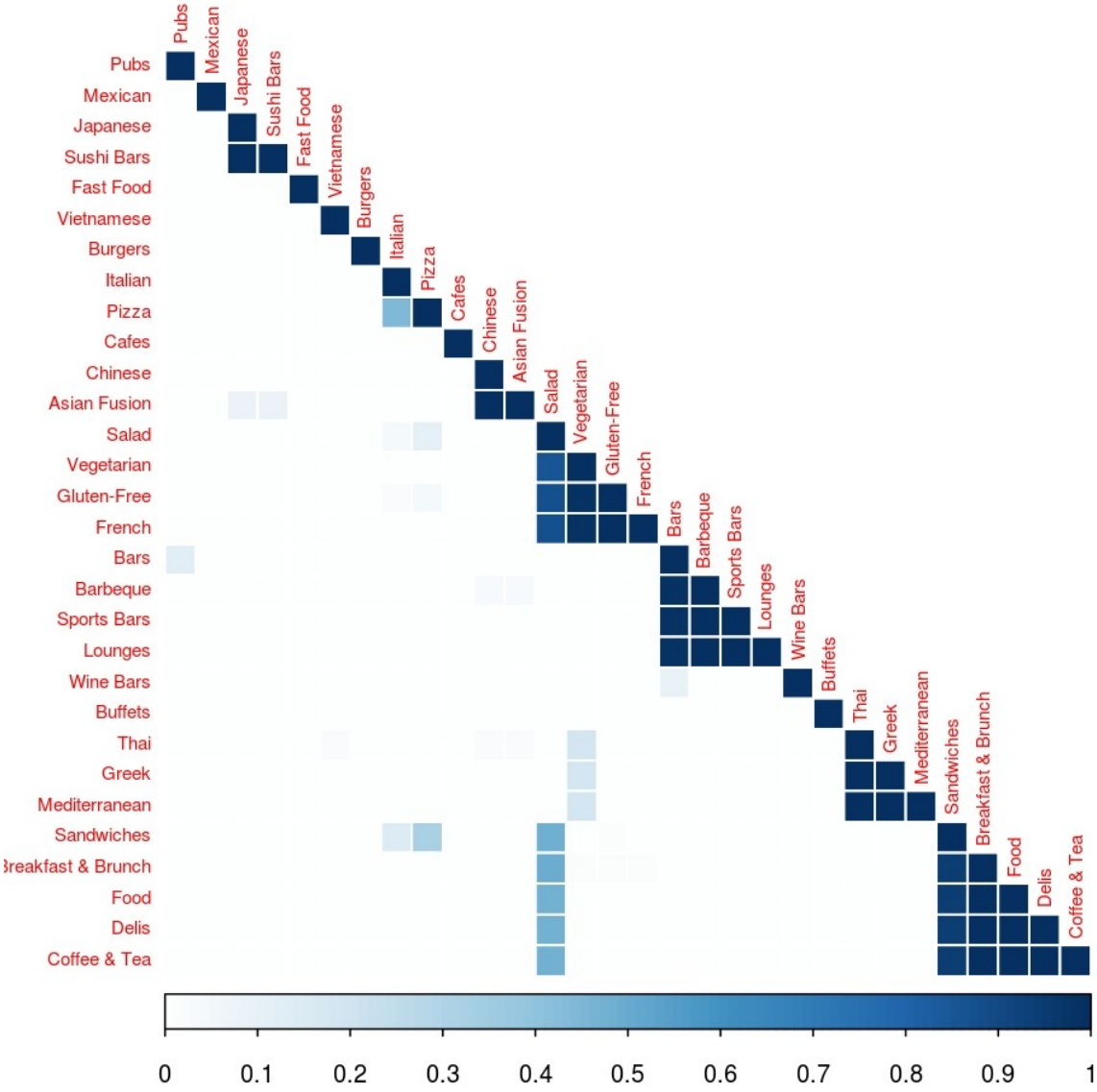

Fig. Clustered Cuisine Map obtained with CountVectorizer from scikit-learn.

Analysis: It is easier to notice the similar cuisine in this clustered map. (Italian, Pizza), (Chinese, Asian Fusion), (Sports Bars, Wine Bars, Lounges) make sensible clusters. However, as pointed out earlier, these are just strong correlation and no weak correlations are picked up using just the token counts in documents.

Task 2.2 Improving the Cuisine Map

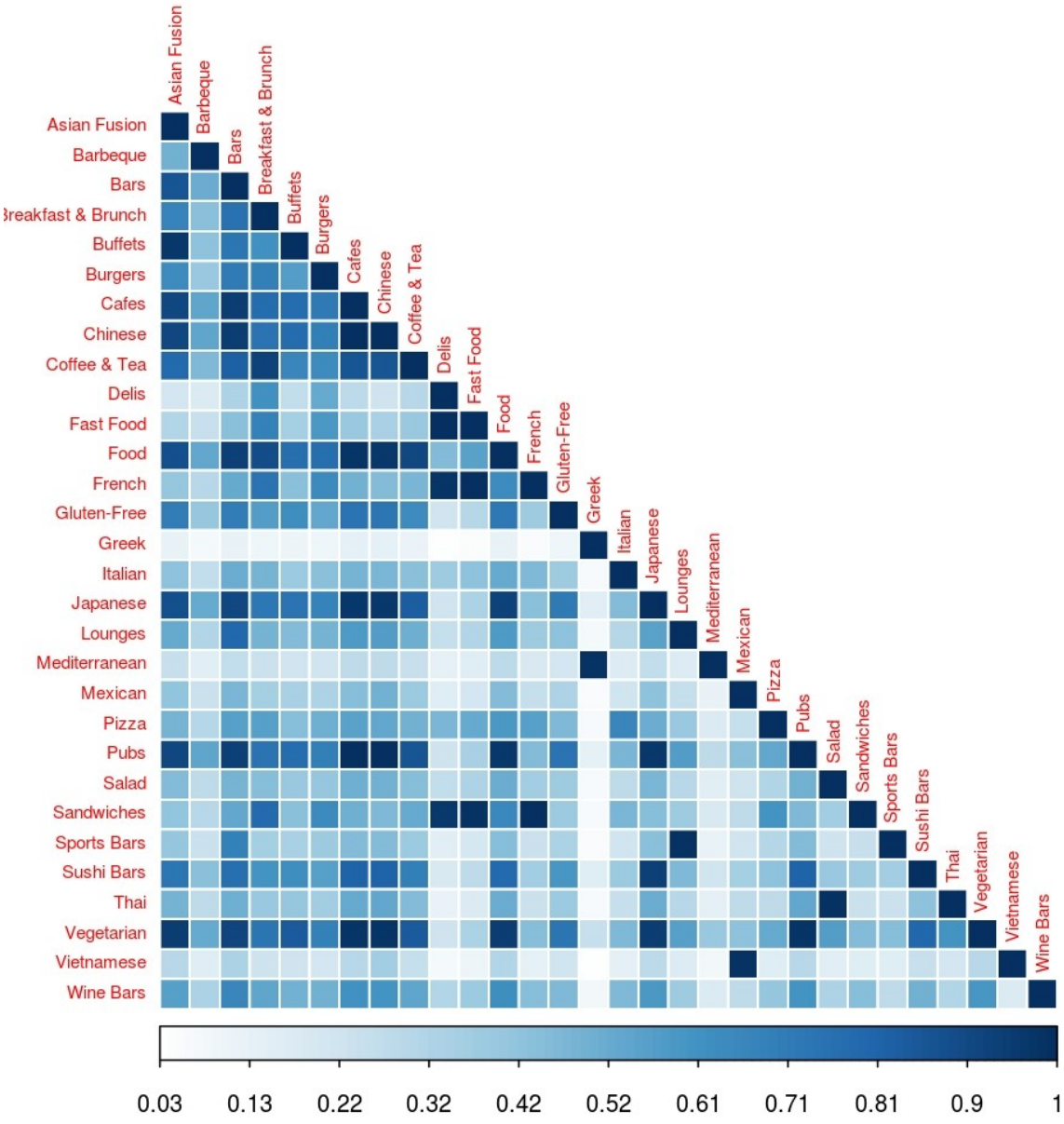

Fig. Cusine Map with no idf, but an Lda topic model applied.

Analysis: I used the TfIdfVectorizer from scikit-learn, without applying idf weighting for this cuisine map. The results are slightly better than using a CountVectorizer but still throw up a lot of false positives (like fairly strong correlation beteen Chinese and Burgers, Chinese and Pubs). In the absence of idf weighting, a lot of common words occuring across the descriptions of these cuisines must be the cause behind these false positives.

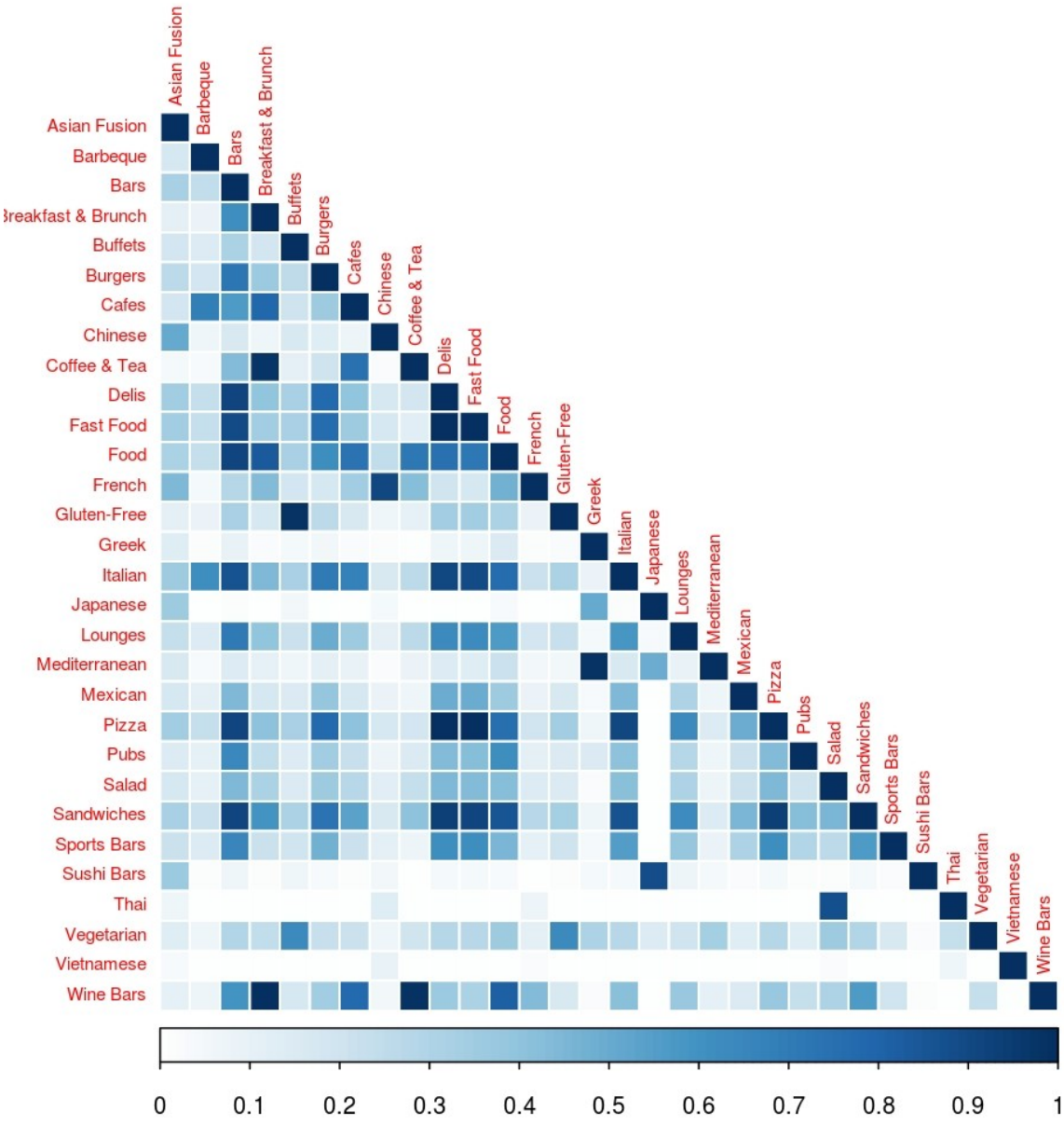

Fig. Cusine Map with idf weighting and a Lda topic model applied.

Analysis: Applying idf weighting (using TfIdfVectorizer with idf weighting) and applying the Lda topic model gives pretty decent results and gets rid of a lot of false positives we saw with no idf. There are fairly strong and genuine correlations like Sushi Bars and Japanese, Coffee & Tea and Breakfast & Brunch etc. Clustering this cuisine map also makes visulization easier as illustrated in the following section.

Task 2.3 Incorporating Clustering in Cuisine Map

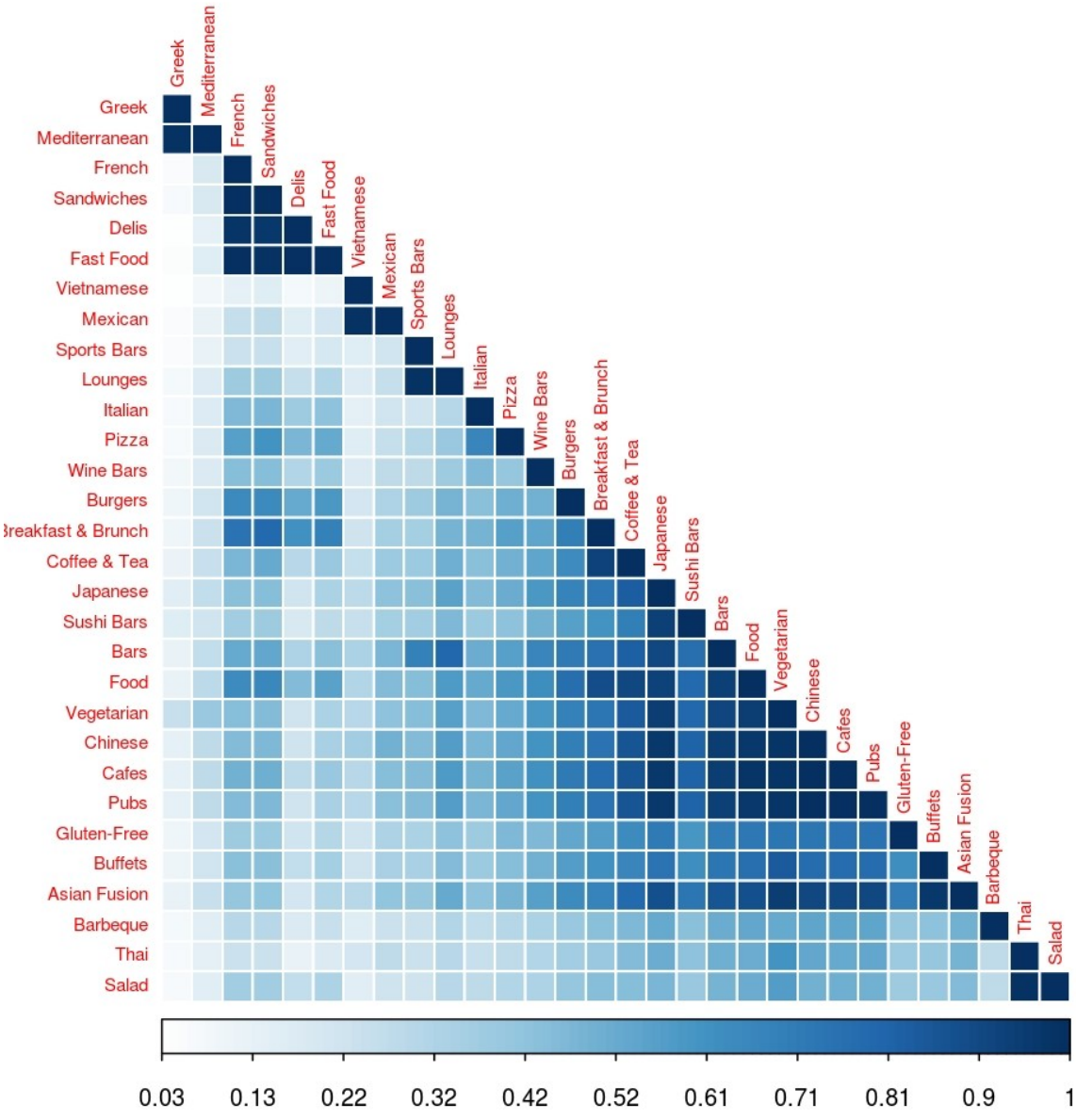

Fig. Clustering with no idf weighting

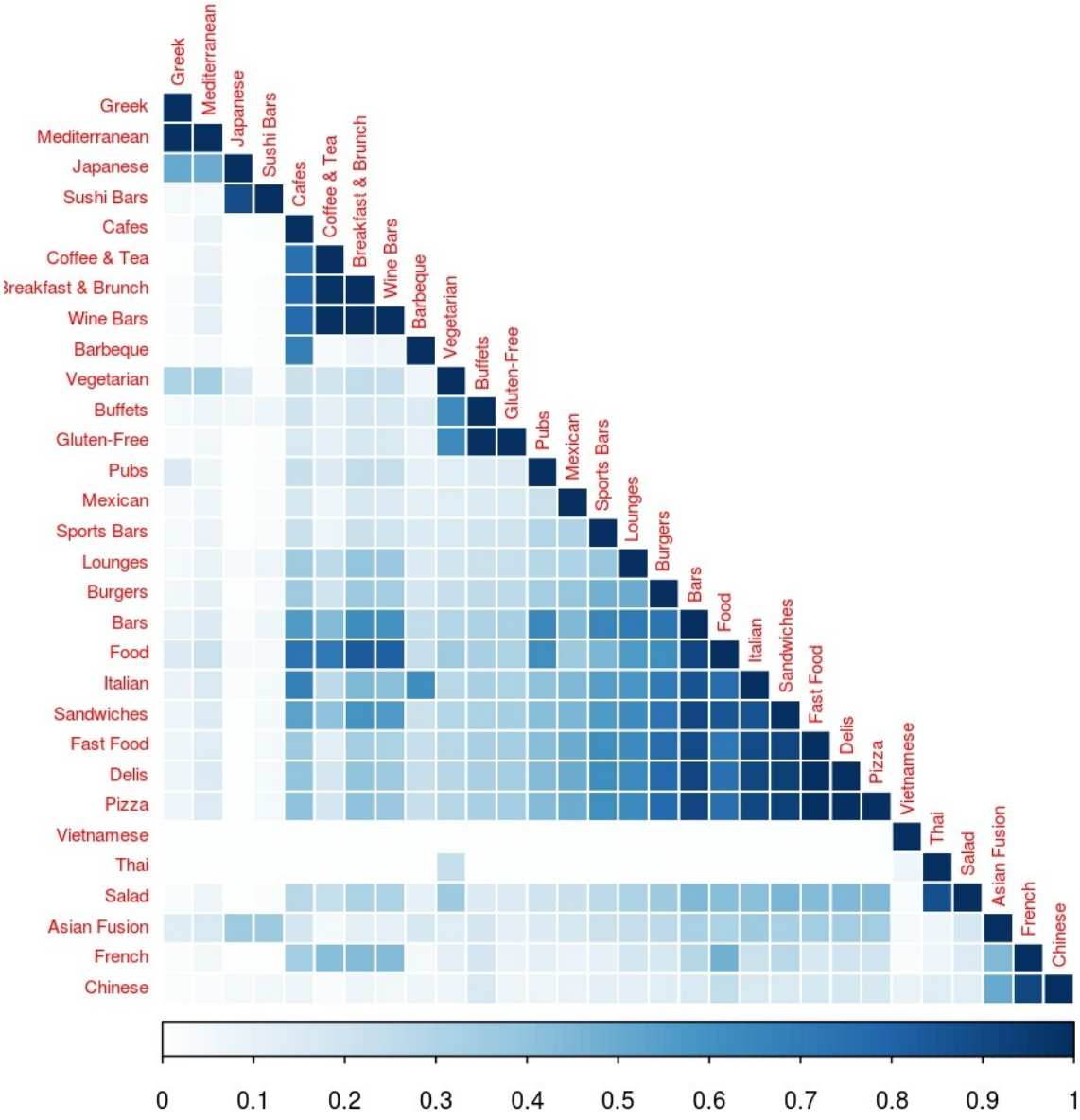

Fig. Clustering with idf weighting

Analysis: Heirarchical clustering has been used to cluster the cuisine map obtained by using the TfIdfVectorizer from scikit-learn with idf weighting and a LDA topic model applied to the 30 most reviewed cuisines. (Greek and Mediterranean), (Sports Bar, Burgers, Lounges, Bars), (Cafes, Coffee & Tea, Breakfast & Brunch) make up pretty sensible clusters. However, there is still some scope for improvement as Thai, Salad, Asian Fusion and Chinese (which are understandably close together) are also closely clustered with French in the visualization.